

# 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture

## ISCA 2016

### Table of Contents

Message from the General Chairs.....	xii
Message from the Program Chair.....	xiii
Organizing Committee.....	xiv
Program Committee.....	xv
External Reviewers.....	xvii
Sponsors.....	xx

---

### Session 1A: Neural Networks I

Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing .....	1
<i>Jorge Albericio, Patrick Judd, Tayler Hetherington, Tor Aamodt, Natalie Enright Jerger, and Andreas Moshovos</i>	
ISAAC: A Convolutional Neural Network Accelerator with In-Situ Analog Arithmetic in Crossbars .....	14
<i>Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R. Stanley Williams, and Vivek Srikumar</i>	
PRIME: A Novel Processing-in-Memory Architecture for Neural Network Computation in ReRAM-Based Main Memory .....	27
<i>Ping Chi, Shuangchen Li, Cong Xu, Tao Zhang, Jishen Zhao, Yongpan Liu, Yu Wang, and Yuan Xie</i>	

### Session 1B: Heterogeneous Architecture/ Approximate Computing

Asymmetry-Aware Work-Stealing Runtimes .....	40
<i>Christopher Torng, Moyang Wang, and Christopher Batten</i>	

Morpheus: Creating Application Objects Efficiently for Heterogeneous Computing .....	53
<i>Hung-Wei Tseng, Qianchen Zhao, Yuxiao Zhou, Mark Gahagan, and Steven Swanson</i>	
Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration .....	66
<i>Divya Mahajan, Amir Yazdanbaksh, Jongse Park, Bradley Thwaites, and Hadi Esmaeilzadeh</i>	

## Session 2A: Caches

Back to the Future: Leveraging Belady's Algorithm for Improved Cache Replacement .....	78
<i>Akanksha Jain and Calvin Lin</i>	
Efficient Synonym Filtering and Scalable Delayed Translation for Hybrid Virtual Caching .....	90
<i>Chang Hyun Park, Taekyung Heo, and Jaehyuk Huh</i>	
LAP: Loop-Block Aware Inclusion Properties for Energy-Efficient Asymmetric Last Level Caches .....	103
<i>Hsiang-Yun Cheng, Jishen Zhao, Jack Sampson, Mary Jane Irwin, Aamer Jaleel, Yu Lu, and Yuan Xie</i>	

## Session 2B: Hardware Design

Automatic Generation of Efficient Accelerators for Reconfigurable Hardware .....	115
<i>David Koeplinger, Raghu Prabhakar, Yaqi Zhang, Christina Delimitrou, Christos Kozyrakis, and Kunle Olukotun</i>	
Strober: Fast and Accurate Sample-Based Energy Simulation for Arbitrary RTL .....	128
<i>Donggyu Kim, Adam Izraelevitz, Christopher Celio, Hokeun Kim, Brian Zimmer, Yunsup Lee, Jonathan Bachrach, and Krste Asanovic</i>	
PowerChop: Identifying and Managing Non-critical Units in Hybrid Processor Architectures .....	140
<i>Michael A. Laurenzano, Yunqi Zhang, Jiang Chen, Lingjia Tang, and Jason Mars</i>	

## Session 3A: Accelerators

Biscuit: A Framework for Near-Data Processing of Big Data Workloads .....	153
<i>Boncheol Gu, Andre S. Yoon, Duck-Ho Bae, Insoon Jo, Jinyoung Lee, Jonghyun Yoon, Jeong-Uk Kang, Moonsang Kwon, Chanho Yoon, Sangyeun Cho, Jaeheon Jeong, and Duckhyun Chang</i>	
Energy Efficient Architecture for Graph Analytics Accelerators .....	166
<i>Muhammet Mustafa Ozdal, Serif Yesil, Taemin Kim, Andrey Ayupov, John Greth, Steven Burns, and Ozcan Ozturk</i>	

ASIC Clouds: Specializing the Datacenter .....	178
<i>Ikuo Magaki, Moein Khazraee, Luis Vega Gutierrez, and Michael Bedford Taylor</i>	

## Session 3B: GPU I

APRES: Improving Cache Efficiency by Exploiting Load Characteristics on GPUs .....	191
<i>Yunho Oh, Keunsoo Kim, Myung Kuk Yoon, Jong Hyun Park, Yongjun Park, Won Woo Ro, and Murali Annavaram</i>	
Transparent Offloading and Mapping (TOM): Enabling Programmer-Transparent Near-Data Processing in GPU Systems .....	204
<i>Kevin Hsieh, Eiman Ebrahim, Gwangsun Kim, Niladrish Chatterjee, Mike O'Connor, Nandita Vijaykumar, Onur Mutlu, and Stephen W. Keckler</i>	
Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming .....	217
<i>Chang Hyun Park, Taekyung Heo, and Jaehyuk Huh</i>	
Warped-Slicer: Efficient Intra-SM Slicing through Dynamic Resource Partitioning for GPU Multiprogramming .....	230
<i>Qiumin Xu, Hyeran Jeon, Keunsoo Kim, Won Woo Ro, and Murali Annavaram</i>	

## Session 4A: Neural Networks II

EIE: Efficient Inference Engine on Compressed Deep Neural Network .....	243
<i>Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark A. Horowitz, and William J. Dally</i>	
RedEye: Analog ConvNet Image Sensor Architecture for Continuous Mobile Vision .....	255
<i>Robert LiKamWa, Yunhui Hou, Yuan Gao, Mia Polansky, and Lin Zhong</i>	
Minerva: Enabling Low-Power, Highly-Accurate Deep Neural Network Accelerators .....	267
<i>Brandon Reagen, Paul Whatmough, Robert Adolf, Saketh Rama, Hyunkwang Lee, Sae Kyu Lee, José Miguel Hernández-Lobato, Gu-Yeon Wei, and David Brooks</i>	

## Session 4B: NoC/Virtualization

Opportunistic Competition Overhead Reduction for Expediting Critical Section in NoC Based CMPs .....	279
<i>Yuan Yao and Zhonghai Lu</i>	
Short-Circuit Dispatch: Accelerating Virtual Machine Interpreters on Embedded Processors .....	291
<i>Channoh Kim, Sungmin Kim, Hyeon Gyu Cho, Dooyoung Kim, Jaehyeok Kim, Young H. Oh, Hakbeom Jang, and Jae W. Lee</i>	

ARM Virtualization: Performance and Architectural Implications .....	304
<i>Christoffer Dall, Shih-Wei Li, Jin Tack Lim, Jason Nieh, and Georgios Koloventzos</i>	

## **Session 5A: Cache/Memory Compression**

Base-Victim Compression: An Opportunistic Cache Compression Architecture .....	317
<i>Jayesh Gaur, Alaa R. Alameldeen, and Sreenivas Subramoney</i>	
Bit-Plane Compression: Transforming Data for Better Compression in Many-Core Architectures .....	329
<i>Jungrae Kim, Michael Sullivan, Esha Choukse, and Mattan Erez</i>	

## **Session 5B: Reliability I**

XED: Exposing On-Die Error Detection Information for Strong Memory Reliability .....	341
<i>Prashant J. Nair, Vilas Sridharan, and Moinuddin K. Qureshi</i>	
Production-Run Software Failure Diagnosis via Adaptive Communication Tracking .....	354
<i>Mohammad Mejbah UI Alam and Abdullah Muzahid</i>	

## **Session 6: Neural Networks III**

Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks .....	367
<i>Yu-Hsin Chen, Joel Emer, and Vivienne Sze</i>	
Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory .....	380
<i>Duckhwan Kim, Jaeha Kung, Sek Chai, Sudhakar Yalamanchili, and Saibal Mukhopadhyay</i>	
Cambricon: An Instruction Set Architecture for Neural Networks .....	393
<i>Shaoli Liu, Zidong Du, Jinhua Tao, Dong Han, Tao Luo, Yuan Xie, Yunji Chen, and Tianshi Chen</i>	

## **Session 7A: Micro Architecture**

Decoupling Loads for Nano-Instruction Set Computers .....	406
<i>Ziqiang Huang, Andrew D. Hilton, and Benjamin C. Lee</i>	
Future Vector Microprocessor Extensions for Data Aggregations .....	418
<i>Timothy Hayes, Oscar Palomar, Osman Unsal, Adrian Cristal, and Mateo Valero</i>	
Efficiently Scaling Out-of-Order Cores for Simultaneous Multithreading .....	431
<i>Faissal M. Sleiman and Thomas F. Wenisch</i>	
Accelerating Dependent Cache Misses with an Enhanced Memory Controller .....	444
<i>Milad Hashemi, Khubaib, Eiman Ebrahimi, Onur Mutlu, and Yale N. Patt</i>	

## Session 7B: Datacenter

Treadmill: Attributing the Source of Tail Latency through Precise Load Testing and Statistical Inference .....	456
<i>Yunqi Zhang, David Meisner, Jason Mars, and Lingjia Tang</i>	
Dynamo: Facebook's Data Center-Wide Power Management System .....	469
<i>Qiang Wu, Qingyuan Deng, Lakshmi Ganesh, Chang-Hong Hsu, Yun Jin, Sanjeev Kumar, Bin Li, Justin Meza, and Yee Jiun Song</i>	
Peak Efficiency Aware Scheduling for Highly Energy Proportional Servers .....	481
<i>Daniel Wong</i>	
Power Attack Defense: Securing Battery-Backed Data Centers .....	493
<i>Chao Li, Zhenhua Wang, Xiaofeng Hou, Haopeng Chen, Xiaoyao Liang, and Minyi Guo</i>	

## Session 8A: Memory I

DRAF: A Low-Power DRAM-Based Reconfigurable Acceleration Fabric .....	506
<i>Mingyu Gao, Christina Delimitrou, Dimin Niu, Krishna T. Malladi, Hongzhong Zheng, Bob Brennan, and Christos Kozyrakis</i>	
Mellow Writes: Extending Lifetime in Resistive Memories through Selective Slow Write Backs .....	519
<i>Lunkai Zhang, Brian Neely, Diana Franklin, Dmitri Strukov, Yuan Xie, and Frederic T. Chong</i>	
MITTS: Memory Inter-arrival Time Traffic Shaping .....	532
<i>Yanqi Zhou and David Wentzlaff</i>	

## Session 8B: Emerging Architectures

The Anytime Automaton .....	545
<i>Joshua San Miguel and Natalie Enright Jerger</i>	
Accelerating Markov Random Field Inference Using Molecular Optical Gibbs Sampling Units .....	558
<i>Siyang Wang, Xiangyu Zhang, Yuxuan Li, Ramin Bashizade, Song Yang, Chris Dwyer, and Alvin R. Lebeck</i>	
Evaluation of an Analog Accelerator for Linear Algebra .....	570
<i>Yipeng Huang, Ning Guo, Mingoo Seok, Yannis Tsividis, and Simha Sethumadhavan</i>	

## Session 9A: GPU II

LaPerm: Locality Aware Scheduler for Dynamic Parallelism on GPUs .....	583
<i>Jin Wang, Norm Rubin, Albert Sidelnik, and Sudhakar Yalamanchili</i>	
ActivePointers: A Case for Software Address Translation on GPUs .....	596
<i>Sagi Shahar, Shai Bergman, and Mark Silberstein</i>	
Virtual Thread: Maximizing Thread-Level Parallelism beyond GPU Scheduling Limit .....	609
<i>Myung Kuk Yoon, Keunsoo Kim, Sangpil Lee, Won Woo Ro, and Murali Annavaram</i>	

## Session 9B: Reliability II

All-Inclusive ECC: Thorough End-to-End Protection for Reliable Computer Memory .....	622
<i>Jungrae Kim, Michael Sullivan, Sangkug Lym, and Mattan Erez</i>	
Rescuing Uncorrectable Fault Patterns in On-Chip Memories through Error Pattern Transformation .....	634
<i>Henry Duwe, Xun Jian, Daniel Petrisko, and Rakesh Kumar</i>	
RelaxFault Memory Repair .....	645
<i>Dong Wan Kim and Mattan Erez</i>	

## Session 10A: Energy Efficient Computing

Using Multiple Input, Multiple Output Formal Control to Maximize Resource Efficiency in Architectures .....	658
<i>Raghavendra Pradyumna Pothukuchi, Amin Ansari, Petros Voulgaris, and Josep Torrellas</i>	
Exploiting Dynamic Timing Slack for Energy Efficiency in Ultra-Low-Power Embedded Systems .....	671
<i>Hari Cherupalli, Rakesh Kumar, and John Sartori</i>	
CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture .....	682
<i>Yanqi Zhou, Henry Hoffmann, and David Wentzlaff</i>	

## Session 10B: Memory II

Boosting Access Parallelism to PCM-Based Main Memory .....	695
<i>Mohammad Arjomand, Mahmut T. Kandemir, Anand Sivasubramaniam, and Chita R. Das</i>	
Agile Paging: Exceeding the Best of Nested and Shadow Paging .....	707
<i>Jayneel Gandhi, Mark D. Hill, and Michael M. Swift</i>	

Energy Efficient Data Encoding in DRAM Channels Exploiting Data Value Similarity .....	719
<i>Hoseok Seol, Wongyu Shin, Jaemin Jang, Jungwhan Choi, Jinwoong Suh, and Lee-Sup Kim</i>	
<b>Author Index</b> .....	731